Advances in Applied Mathematics 应用数学进展, 2022, 11(11), 8222-8230
Published Online November 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.1111870

Hans汉斯

基于股吧评论的投资者情绪与股市波动研究

——以九安医疗为例

叶陆平

上海工程技术大学数理与统计学院，上海

收稿日期：2022年10月23日；录用日期：2022年11月18日；发布日期：2022年11月29日

摘　要

以新冠疫情爆发为研究背景，针对东方财富股吧中大量能反映股吧舆情的文本数据，本文选取九安医疗2022年1月15日到2022年10月15日的股吧评论数据，对比多种情感分析模型，选用Bert + Bi-LSTM模型对投资者情绪进行分类。样本期内积极情感占比为41%，消极情感占比为59%，利用构建的投资者情感指数与股价涨跌幅进行多元线性回归模型，揭示股吧中投资者情绪变动与股市波动有关系，并显示短期内有正向预测效果，长期内有负向预测效果。但本文实证的样本期较短，选取的股吧数据平台较单一，可能对实验结果有一定影响。

关键词

文本情感分析，Bert + Bi-LSTM，投资者情绪指数，涨跌幅

Research on Investor Sentiment and Stock Market Volatility Based on Stock Bar Comments

—Taking Jiu'an Medical as an Example

Luping Ye

School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai

Received: Oct. 23rd, 2022; accepted: Nov. 18th, 2022; published: Nov. 29th, 2022

Abstract

Taking the outbreak of the COVID-19 as the research background, and aiming at a large number of

文章引用: 叶陆平. 基于股吧评论的投资者情绪与股市波动研究[J]. 应用数学进展, 2022, 11(11): 8222-8230.
DOI: 10.12677/aam.2022.1111870

text data in the Oriental Fortune Stock Bar that can reflect the public opinion of the stock bar, this paper selects the review data of the stock bar of Jiu'an Medical from January 15, 2022 to October 15, 2022, compares a variety of emotion analysis models, and selects Bert + Bi-LSTM model to classify investor sentiment. During the sample period, the proportion of positive emotions was 41%, and the proportion of negative emotions was 59%. Using the constructed investor sentiment index and the stock price rise and fall, we conducted a multiple linear regression model to reveal the relationship between investor sentiment changes in the stock bar and stock market volatility, and showed that there was a positive prediction effect in the short term and a negative prediction effect in the long term. However, the sample period of this empirical study is short, and the stock bar data platform selected is single, which may have a certain impact on the experimental results.

Keywords

Text Sentiment Analysis, Bert + Bi-LSTM, Investor Sentiment Index, Rise and Fall

1. 引言

自 2019 年底开始，新冠疫情在全球范围内大幅度传播，世界经济发展受到显著影响，实体经济大规模停摆，人们对经济的悲观预期加重，社会舆论和国家政策也对经济产生影响。而随着网络的普及，网民们在各社交平台纷纷表达自己的意见，在股市中主要体现在投资者在“东方财富股吧”、“新浪微博”平台上发布自己的投资情绪。行为金融理论提出一个“非理性人”假设，即投资者并非都是理性的，在进行投资过程中往往会受到他人的观点影响[1]，产生羊群效应等。

随着社交媒体的发展，投资者在线参与股吧评论的意愿逐渐强烈，东方财富股吧每天都产生大量的投资者文本评论，并且其具有传播速度快、影响力大、实时性强的特点，在股吧中进行评论的人员不仅有专业投资者，还有大量散户投资者，因此能够反映出整个社会对股市的及时性看法。通过对股吧评论进行详细文本情感分析，分析投资者在一定时期的投资倾向有助于提取和量化市场情感，对了解股市情况，提出未来的投资建议有重要意义。

2. 相关工作

在文本分析领域的研究中，情感分析研究是其中重要的研究内容，赵妍妍等[2]分别从情感信息抽取、情感信息分类和情感信息的检索与归纳三个角度详细的介绍了情感分析的研究现状，为后续研究者提供一些启示。与此同时，文本情感分析研究主要分为三方面：基于情感词典的情感分析方法、基于传统机器学习的情感分析方法和基于深度学习的情感分析方法[3]。Baccianella 等[4]提出 SentiWordNet 情感词典法，该词典利用 WordNet 把含义一致或相近的词语汇集到一起，并手动赋予词语的情感极性分数，而中文情感词典与此不同，中文词典是将褒义词、中性词、贬义词进行划分，利用情感词典进行情感分析的算法主要有 SnowNLP 和 BosonNLP。然而，在此基础上建立的情感词典具有语言单一性的缺点，栗雨晴等人[5]在总结前人的基础上，证实了利用双语词典进行情感分析准确率更高，构建了基于双语词典的情感分析模型。但是研究中并没考虑同一词语在不同领域分类含义不同的问题，因此 Cai 等[6]提出一种可以在特定领域使用的情感词典，并通过实验证明将支持向量机算法和梯度下降树算法进行集成得到的分

类效果更好，但在研究中并未考虑中文情感词典中词汇量较少，口语化表达、网络热词、谐音梗等问题，赵妍妍等[7]从微博平台出发，基于微博平台挖掘网络热词，构建由微博热词相关的情感词典，该词典在微博情感分析中性能提升 1.13%。但随着互联网时代的发展，信息传播速度加快，网络热词出现的速度飞快，情感词典具有时效性，对于许多后来出现的网络特殊用语、谐音梗、表情符号，会导致分类效果变差，现有的情感词典随着时间的流逝性能会逐渐变差，需要不断补充新词汇。

机器学习是多领域交叉学科，重点研究对象是人工智能，主要功能是利用以往数据和经验，优化算法性能。基于机器学习的情感分析方法是目前挖掘社交网络文本信息的重要方法，主要分为有监督学习、无监督学习和半监督学习[8]。在有监督学习中，通过带标签的样本进行训练，从顶向下进行训练，不断微调得到最优参数，对样本的依赖性较高，需要大量已做标签的样本，常用算法包括：朴素贝叶斯、支持向量机和逻辑回归等[3]。无监督学习是从底层开始，从下往上训练各层参数，常用算法是 K 均值算法和 Apriori 算法，在情感分类中较少使用。半监督学习输入的数据由不带标签数据和带标签数据组成，是无监督学习的延伸，与有监督学习相比，减少了对大量有标签样本的依赖，降低了标注成本。国内外很多学者通过机器学习算法对文本情感进行分类，Pang 等[9]利用支持向量机、最大熵分类、朴素贝叶斯算法构建情感分类模型，对电影评论进行情感分析，实验表明，支持向量机分类效果最好，达到 82.9%。2018 年 Google 公司提出 Bert 模型[10]之后，基于深度学习的文本情感分析方法得到广泛应用，预测精度也取得大幅度提升。

目前，基于新冠疫情背景下，我国学者以股吧投资者文本情感分析为出发点研究股票动态的研究较少，本文基于多种情感分析模型，选出适合股吧文本特征的模型进行文本情感分析，有利于丰富此部分研究。

3. 股吧评论文本舆情分析

为保证股吧文本分析的完整性，使读者对疫情期间股吧舆情分析步骤有一个基本的了解，本文从东方财富股吧中爬取文本数据，进行一系列的文本分析，得到样本期间内股吧评论的热点问题，并且对股吧中每日得到的股评信息进行情感分类，得到每一交易日的股吧投资者情绪倾向，为基于投资者情绪的股市波动关联分析提供基础。

3.1. 基于时间序列的热度分析

为了更直观的分析股吧中投资者的讨论热度，本文通过“八爪鱼”爬虫软件对九安医疗股吧进行文本评论爬取，选取 2022 年 1 月 15 日到 2022 年 10 月 15 日的文本评论数据，经过重复数据去重及删除无关信息共得到 378,787 条文本数据。同时，为表现出投资者在不同时期时投资热情不同，本研究对每日爬取的帖子数量进行时间序列分析，如图 1 所示。

由图 1，在选取的研究区间内，共经过两次热度高峰期，分别为 2022 年 1 月 18 日和 2022 年 4 月 15 日，并且可以清晰的观察到，该股吧股评数据量呈周期性，以“周”为单位进行波动。

3.2. 基于股吧文本的情感分析方法

本文针对股吧评论进行文本分析，重点是分析股吧中投资者的情感倾向，通过投资者倾向研究股吧舆情，为投资者进行股票量化研究提供参考。监管部门和上市公司也可以根据股吧舆情及时采取干预措施，矫正舆论方向。情感分析是自然语言处理领域的一种研究较成熟的分析方法，本文选取 6 种常用情感分析算法，在爬取的数据集中随机抽取 10,000 条文本评论进行人工标注，分为积极和消极两类。训练集和测试集按照 7:3 划分，模型分类结果如表 1 所示。

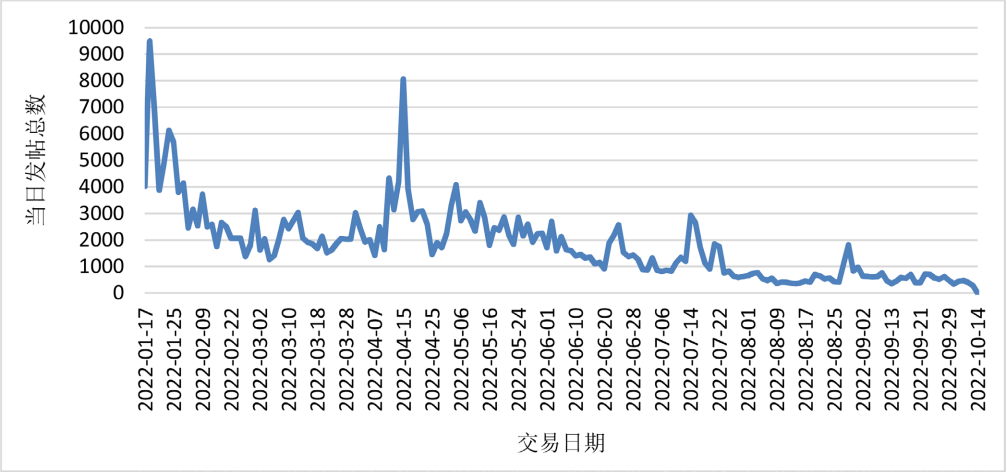

Figure 1. Change of daily posts
图 1. 日发帖量变动情况

Table 1. Comparison of emotion classification results
表 1. 情感分类结果对比

模型名称	accuracy	precision	recall	f1-score
词典法	0.64	0.68	0.65	0.66
随机森林	0.75	0.77	0.75	0.76
决策树	0.70	0.74	0.70	0.71
支持向量机	0.78	0.81	0.78	0.79
朴素贝叶斯类	0.79	0.80	0.79	0.79
Bert + Bi-LSTM	0.85	0.83	0.85	0.84

综合来看，Bert + Bi-LSTM 模型的分类结果在本文数据集上表现最好，因此本文选用 Bert + Bi-LSTM 模型对爬取的股吧评论进行情感分析。BERT 是一个热门的基于预训练的语言表征模型[10]。该模型采用新的 Masked Language Model (MLM)，该模型通过掩盖掉句子中的一些单词，然后通过上下文来预测该单词，从而解决文本分类模型上下文语境缺失的问题。本质来说，MLM 主要目标是去判断或计算词与词之间潜在的语义关系，该模型进行训练时，有 15%的单词会被随机掩盖掉。在这 15%被掩盖掉的单词中，为了提高模型的泛化能力，有 80%会被直接替换成标志符，10%会被替换成其他任意的单词，10%会被保留原始的单词。本文将 Bert 做为嵌入层提取特征，然后传入 Bi-LSTM，最后使用全连接层输出分类，得到本文所使用的二分类模型。最终得到情感分类结果如图 2，负面情绪占比大于正面情绪，这也表明负面情绪更易传播。

4. 基于投资者情绪的股市波动关联分析

本节所用股票相关数据从国泰安数据库取得，样本周期是 2022 年 1 月 19 日到 2022 年 10 月 14 日，采用日频率数据。

4.1. 情绪指标构建

本文利用 Antweiler 和 Frank [11]提出的情绪指标构建方法，令 $Sentiment_t$ 代表股票在第 t 日时股吧中投资者的情绪指数，其定义如下：

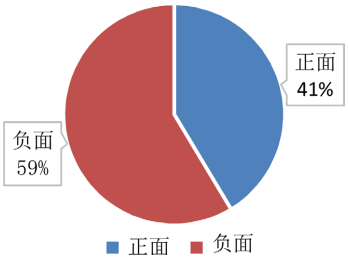

Figure 2. Emotion ratio chart
图 2. 情绪占比图

$$\text{Sentiment}_t = \ln\frac{1+\text{Sen}_t^{\text{看涨}}}{1+\text{Sen}_t^{\text{看跌}}} \tag{1}$$

其中，$\text{Sen}_t^{\text{看涨}}$ 代表第 t 日股吧中看涨帖子的数量，$\text{Sen}_t^{\text{看跌}}$ 代表第 t 日股吧中看跌帖子的数量，由(1)可知 Sentiment_t 为正值时，股吧中情绪偏向于看涨，表明股吧评论中更多投资者预测股价会涨。反之，当 Sentiment_t 为负值时，股吧中情绪偏向于看跌，表明股吧中更多的投资者倾向于卖出观点，股价大概率会下跌。

4.2. 情绪指标的描述性统计

本节通过投资者情绪指标的构建得到如下直方图 3，可以看出大部分投资者表现为保守主义，对股价持有悲观预期。

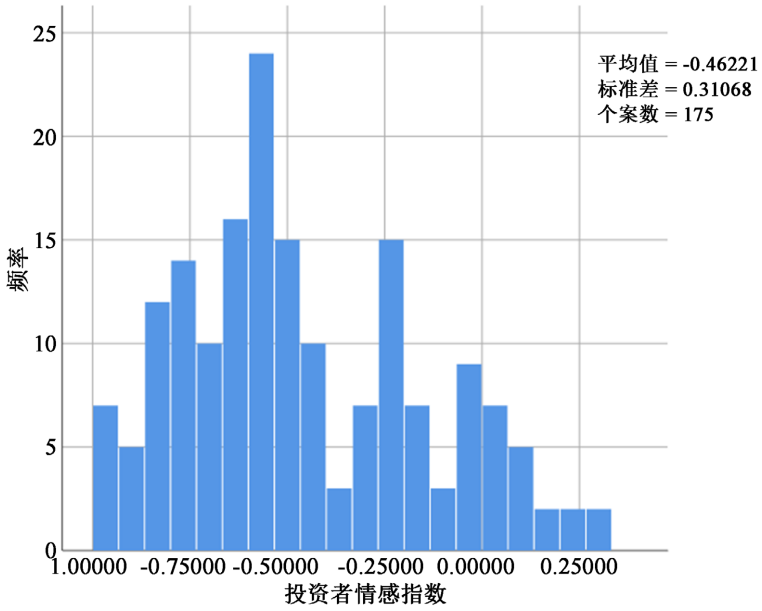

Figure 3. Histogram of investor sentiment index
图 3. 投资者情绪指数直方图

同时，本文结合箱线图 4 也可看出，在取样区间内投资者情绪指数的中位数都小于 0，即样本内的股吧评论大多为看跌评论，投资者大多呈悲观情绪。同时，这也可以反映出投资者更喜欢在处于悲观状态时发表自己对股票的见解，而且悲观情绪的传播速度比乐观情绪更快，这种现象会使股吧中其他参与者受到情绪干扰，影响自己的理性决策，股票的交易波动加剧，形成股吧舆情。

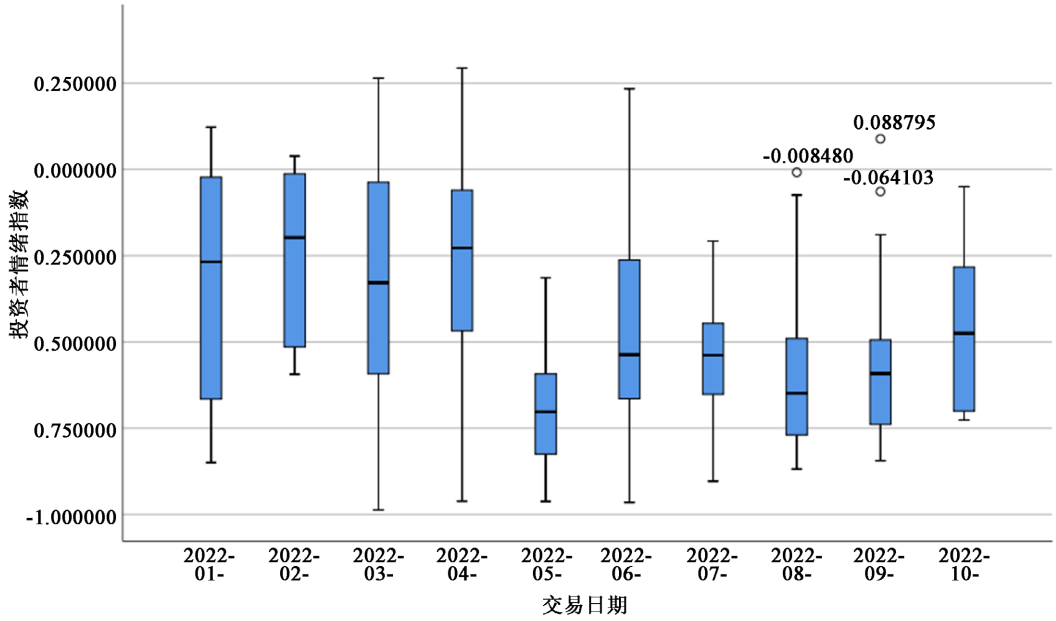

Figure 4. Online chart of investor sentiment index
图 4. 投资者情绪指数线箱图

4.3. 投资者情绪指数与股票涨跌幅波动分析

本节采用股票涨跌幅衡量股票市场的波动情况，一只股票的涨跌幅度是以每日的收盘价与前一日的收盘价相比较，正值为涨，负值为跌。令 CR_t 代表第 t 日股票的涨跌幅，$Close_t$ 代表第 t 日的收盘价，$Close_{t-1}$ 代表第 $t-1$ 日的收盘价，则涨跌幅定义如下：

$$CR_t = \frac{Close_t}{Close_{t-1}} - 1 \tag{2}$$

为了验证涨跌幅、发帖总量、情绪指数之间具有相关性，利用皮尔逊相关系数来评估三者之间的相关性，如表 2。皮尔逊相关系数越接近于 1 或者−1 表明两者之间的相关性越强。由统计分析结果可知，发帖总量与涨跌幅相关性为−0.115 < 0.3，相关性较弱，且未通过相关性检验，本文判定为两者无关。情绪指数与涨跌幅相关性为 0.703 > 0.7，具有较强的相关性，因而本文利用情绪指数来研究股价涨跌幅的波动情况。

Table 2. Correlation analysis
表 2. 相关性分析

		涨跌幅	发帖总量	情绪指数
涨跌幅	皮尔逊相关性	1	−0.115	0.703**
	Sig.(双尾)		0.128	0.000
	个案数	175	175	175
发帖总量	皮尔逊相关性	−0.115	1	0.221**
	Sig.(双尾)	0.128		0.003
	个案数	175	175	175
情绪指数	皮尔逊相关性	0.703**	0.221**	1
	Sig.(双尾)	0.000	0.003	
	个案数	175	175	175

**在 0.01 级别(双尾)，相关性显著。

本文采用经典的多元线性回归方法来研究股吧中投资者情绪对股价的影响情况，参考黄玉婷等[12]做法，本文设定投资者情绪指标滞后 2 阶，股价涨跌幅滞后 1 阶进行多元统计建模，并提出如下统计模型：

$$CR_t = \alpha + \theta_1 CR_{t-1} + \theta_2 Sent_t + \theta_3 Sent_{t-1} + \theta_4 Sent_{t-2} + \varepsilon_t \quad (3)$$

Table 3. Multiple linear regression results
表 3. 多元线性回归结果

	系数	t	显著性	容差
α	0.030	5.735	0.000	
CR_{t-1}	−0.070	−0.947	0.095	0.346
$Sent_t$	0.146	18.707	0.000	0.792
$Sent_{t-1}$	0.051	−3.810	0.000	0.267
$Sent_{t-2}$	−0.026	−2.748	0.007	0.540

继而对数据进行统计分析，得到调整后的 $R^2 = 0.670$，表明整个多元线性模型效果较好。由表 3，则多元线性回归模型为：

$$CR_t = 0.03 - 0.07CR_{t-1} + 0.146Sent_t + 0.051Sent_{t-1} + -0.26Sent_{t-2} + \varepsilon_t \quad (4)$$

表 3 中给出了模型的系数及共线性统计结果。根据表中的容差值均大于 0.1 可知，模型没有多重共线性。根据表中各自变量系数表明，东方财富股吧中的投资者评论会影响股价的涨跌幅，对其进行分析研究可以用于预测股价波动情况。θ_1 系数值为−0.07，其显著性为 0.095 < 0.1，通过显著性检验，上一交易日的涨跌幅对下一交易日涨跌幅具有负向预测作用，但其影响较低，假设在 t − 1 日股价下跌，在 t 日将会有买家以低价购入股票，股票在二级市场的流通中价格会快速回升，股价涨跌幅随之减少甚至变为正值。θ_2 为 0.146 且通过显著性检验，表明根据股吧评论构建的投资者情绪指数可正向反映出当日的股价涨跌幅，当投资者情绪为负值时，股价下跌，反之上涨。θ_4 为情绪指数一阶滞后的系数，该系数为 0.051 且通过显著性检验，表明第 t − 1 日投资者情绪越高涨，第 t 日该股价的上涨概率越大，意味着通过对股吧第 t − 1 日进行研究挑选出看涨情绪高的股票，第 t 日买入会更高概率取得收益。反之，第 t − 1 日投资者情绪为负值的股票，第 t 日股价可能也会下跌。θ_4 系数值为−0.026 且其通过显著性检验，这表明从较长期来看，股吧评论的情感倾向对股票涨跌幅具有负向预测作用。该结果可由“羊群效应”解释，当股吧中长期情绪积累形成一种悲观或者乐观舆情时，投资者会放大这种情绪，导致认知偏差，做出错误的投资决策，影响股价涨跌幅，导致收益率降低。本文以九安医疗股票为例，得到样本期内投资者情绪指数与股票涨跌幅折线图，如下图。

由图 5，可以得出投资者情绪指数变动和股票涨跌幅波动幅度基本一致，即研究股吧中投资者的情绪指数对股价波动情况分析具有重要意义。

5. 结论及展望

本文以九安医疗为研究对象，基于东方财富股吧评论的投资者情绪研究其与股市波动之间的关系。首先基于时间序列分析股吧发帖热度，再使用多种情感分析模型对比，选出精度最高的分类模型对东方财富股吧中的投资者股票评论进行情感分类，进而根据分类结果构建投资者情感指数，最后将投资者情

绪指数与股价涨跌幅构建多元线性回归模型。经过一系列实证得出投资者情感情绪与股价波动具有一定关系，综合股吧舆情和股票量化投资角度得到结论如下：

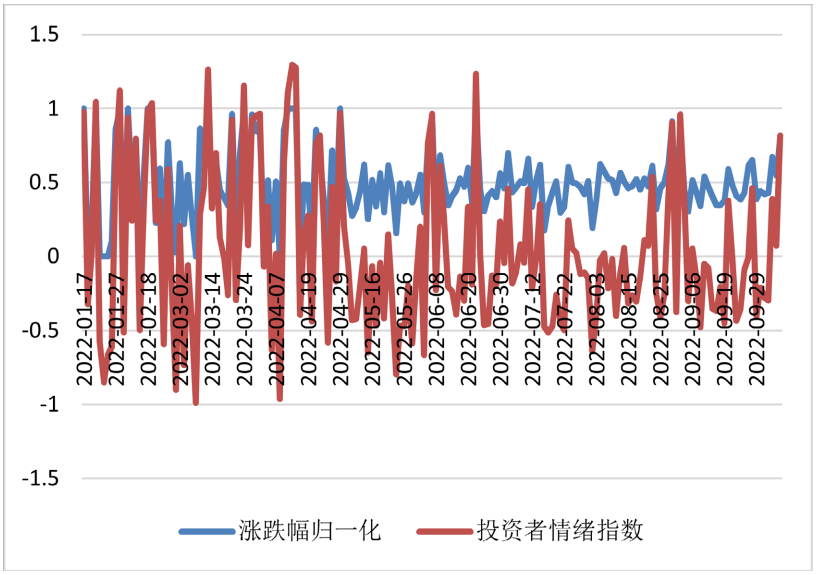

Figure 5. Broken line chart of investor sentiment index and stock fluctuation
图 5. 投资者情绪指数与股票涨跌幅折线图

随着互联网发展，投资者的“羊群效应”现象更加明显，股吧中的乐观情绪或者是悲观情绪更容易被放大，股吧中存在着大量有影响力的评论者，通过反复刷屏、抛出带有主观色彩的“内幕消息”误导其他投资者。这种现象的存在易引发股市舆情，加剧股市波动，导致大量散户投资者被“割韭菜”，应该引起监管部门重视，及时发现此类舆情，采取积极措施进行主动干预。涉及的上市公司也应加强对类似舆情的监测，及时发布公开说明，避免散户投资者被误导，以免股价产生大幅波动。

本文研究结果可表明，投资者情绪可对股价产生影响，在股票量化投资时加入投资者情绪指数或者可代表投资者情绪的其他宏观经济因子，均会提高选股收益。投资者在构建股票投资组合或者进行择时交易时加入情感因子，在适时买入或卖出股票，避免被主力“割韭菜”，而且有助于股市的健康有序发展。

除此之外，本文有许多不足，比如在对东方财富股吧进行情感分类时，分类准确度还需要不断提升，选取的股吧数据平台较单一，理论上应结合多平台数据对比分析才能提高预测的精确率和提高数据的有效性。本文仅对九安医疗一只股票进行文本分析，后续会继续研究多只股票的对比分析。随着网络的发展，对大数据进行处理和分析将成为金融研究的主流，并会对股市健康平稳发展提供支持作用。

参考文献

[1] Hudson, R. and Muradoglu, Y.G. (2020) Personal Routes into Behavioural Finance. *Review of Behavioral Finance*, **12**, 1-9. https://doi.org/10.1108/RBF-12-2019-0176

[2] 赵妍妍, 秦兵, 石秋慧, 刘挺. 大规模情感词典的构建及其在情感分类中的应用[J]. 中文信息学报, 2017, 31(2): 187-193.

[3] 王婷, 杨文忠. 文本情感分析方法研究综述[J]. 计算机工程与应用, 2021, 57(12): 11-24.

[4] Baccianella, S., Esuli, A. and Sebastiani, F. (2010) Sentiwordnet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (*LREC*'10), 2010.

[5] 栗雨晴, 礼欣, 韩煦, 宋丹丹, 廖乐健. 基于双语词典的微博多类情感分析方法[J]. 电子学报, 2016, 44(9): 2068-2073.

[6] Cai, X.H., Liu, P.Y., Wang, Z.H., *et al*. (2016) Fine-Grained Sentiment Analysis Based on Sentiment Disambiguation. 2016 8*th International Conference on Information Technology in Medicine and Education* (*ITME*), Fuzhou, 23-25 December 2016, 557-561. https://doi.org/10.1109/ITME.2016.0132

[7] 赵妍妍, 秦兵, 石秋慧, 刘挺. 大规模情感词典的构建及其在情感分类中的应用[J]. 中文信息学报, 2017, 31(2): 187-193.

[8] 刘宇. 基于网络舆情的量化选股策略实证研究[D]: [硕士学位论文]. 成都: 西南民族大学, 2019. https://doi.org/10.27417/d.cnki.gxnmc.2019.000089

[9] Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs Up? Sentiment Classification Using Machine Learning Techniques. *Proceedings of the ACL*-02 *Conference on Empirical Methods in Natural Language Processing*, **10**, 79-86. https://doi.org/10.3115/1118693.1118704

[10] Tom, Y., Devamanyu, H., Soujanya, P. and Erik, C. (2018) Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Computational Intelligence Magazine*, **13**, 55-75.

[11] Antweiler, W. and Frank, M.Z. (2004) Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance*, **59**, 1259-1294. https://doi.org/10.1111/j.1540-6261.2004.00662.x

[12] 黄雨婷, 宋泽芳, 李元. 基于文本挖掘的股评情绪效应分析[J/OL]. 数理统计与管理: 1-14, 2022-06-25. https://doi.org/10.13860/j.cnki.sltj.20211130-010